开放式情境判断测验的自动化评分*

徐静[1] 骆方[1] 马彦珍[2] 胡路明[3] 田雪涛[1]

([1]北京师范大学心理学部, 北京 100875)

([2]中国基础教育质量监测协同创新中心, 北京 100875)

([3]北京师范大学珠海校区文理学院心理系, 珠海 519085)

摘要 受限于评分成本，开放式情境判断测验难以广泛使用。本研究以教师胜任力测评为例，探索了自动化评分的应用。针对教学中的典型问题场景开发了开放式情境判断测验，收集中小学教师作答文本，采用有监督学习策略分别从文档层面和句子层面应用深度神经网络识别作答类别，卷积神经网络(Convolutional Neural Network, CNN)效果理想，各题评分准确率为70%~88%，与人类评分一致性高，人机评分的相关系数 *r* 为 0.95, 二次加权 Kappa 系数(Quadratic Weighted Kappa, QWK)为 0.82。结果表明，机器评分可以获得稳定的效果，自动化评分研究能够助力于开放式情境判断测验的广泛应用。

关键词 情境判断测验，自动化评分，教师胜任力，开放式测验，机器学习

Automated Scoring of Open-ended Situational Judgment Tests

XU Jing[1], LUO Fang[1] , MA Yanzhen[2], HU Luming[3], TIAN Xuetao[1]

([1]*School of Psychology, Beijing Normal University, Beijing* 100875, *China*)

([2]*Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University, Beijing* 100875, *China*)

([3]*Department of Psychology, School of Arts and Sciences, Beijing Normal University at Zhuhai, Zhuhai 519085 , China*)

Abstract

Situational Judgment Tests (SJTs) have gained popularity for their unique testing content and high face validity. However, traditional SJT formats, particularly those employing multiple-choice (MC) options, have encountered scrutiny due to their susceptibility to test-taking strategies. In contrast, open-ended and constructed response (CR) formats present a propitious means to address this issue. Nevertheless, their extensive adoption encounters hurdles primarily stemming from the financial implications associated with manual scoring. In response to this challenge, we propose an open-ended SJT employing a written-constructed response format for the assessment of teacher competency. This study established a scoring framework leveraging natural language processing

收稿日期: 2022-10-22

* 国家自然科学基金青年科学基金(62207002); 国家自然科学基金面上项目(62377003); 中国博士后科学基金特别资助(站前)(2022TQ0040); 中国博士后科学基金面上资助(2022M720486)。

通信作者: 田雪涛, E-mail: xttian@bnu.edu.cn

(NLP) technology to automate the assessment of response texts, subsequently subjecting the system's validity to rigorous evaluation. The study constructed a comprehensive teacher competency model encompassing four distinct dimensions: student-oriented, problem-solving, emotional intelligence, and achievement motivation. Additionally, an open-ended situational judgment test was developed to gauge teachers' aptitude in addressing typical teaching dilemmas. A dataset comprising responses from 627 primary and secondary school teachers was collected, with manual scoring based on predefined criteria applied to 6,000 response texts from 300 participants. To expedite the scoring process, supervised learning strategies were employed, facilitating the categorization of responses at both the document and sentence levels. Various deep learning models, including the convolutional neural network (CNN), recurrent neural network (RNN), long short-term memory (LSTM), C-LSTM, RNN+attention, and LSTM+attention, were implemented and subsequently compared, thereby assessing the concordance between human and machine scoring. The validity of automatic scoring was also verified.

This study reveals that the open-ended situational judgment test exhibited an impressive Cronbach's alpha coefficient of 0.91 and demonstrated a good fit in the validation factor analysis through the use of Mplus. Criterion-related validity was assessed, revealing significant correlations between test results and various educational facets, including instructional design, classroom evaluation, homework design, job satisfaction, and teaching philosophy. Among the diverse machine scoring models evaluated, CNNs have emerged as the top-performing model, boasting a scoring accuracy ranging from 70% to 88%, coupled with a remarkable degree of consistency with expert scores (r= 0.95, QWK=0.82). The correlation coefficients between human and computer ratings for the four dimensions—student-oriented, problem-solving, emotional intelligence, and achievement motivation—approximated 0.9. Furthermore, the model showcased an elevated level of predictive accuracy when applied to new text datasets, serving as compelling evidence of its robust generalization capabilities.

This study ventured into the realm of automated scoring for open-ended situational judgment tests, employing rigorous psychometric methodologies. To affirm its validity, the study concentrated on a specific facet: the evaluation of teacher competency traits. Fine-grained scoring guidelines were formulated, and state-of-the-art NLP techniques were used for text feature recognition and classification. The primary findings of this investigation can be summarized as follows: (1) Open-ended SJTs can establish precise scoring criteria grounded in crucial behavioral response elements; (2) Sentence-level text classification outperforms document-level classification, with CNNs exhibiting remarkable accuracy in response categorization; and (3) The scoring model consistently delivers robust performance and demonstrates a remarkable degree of alignment with human scoring, thereby hinting at its potential to partially supplant manual scoring procedures.

Keywords situational judgment tests, automated scoring, teacher competency, open-ended tests, machine learning.

1 引言

在人事测评领域，情境判断测验(Situational Judgment Test, SJT)因其测验内容的独特性和较高的表面效度而广为流行，常用于人员选拔与评估。题干通常呈现一系列与工作相关的情境，选项则是若干典型行为反应，要求受测者选择最符合自己实际做法的一项或对选项排序(漆书青, 戴海琦, 2003)。情境判断测验是测量胜任力的良好工具，比面试成本更低，比自陈式量表更生动，在预测工作绩效方面比一般认知能力测验、人格测验表现更佳(Burrus et al., 2012; McDaniel et al., 2007; McDaniel et al., 2011; Oostrom et al., 2012; Slaughter et al., 2014; Weekley & Ployhart, 2005)。

按照开放程度的不同，情境判断测验作答形式总体可分为封闭式(Closed Response Formats)和开放式(Open-ended Formats)。封闭式即传统的多项选择式(Multiple Choice, MC)；开放式即构答反应式(Constructed Response, CR)，题目不呈现选项，被试可自由作答，主要包括书面回答式(Written-constructed)、视听构建式(Audio-visual Constructed)、情景面试(Situational Interview)等。其中，书面回答式要求受测者写出做法；视听构建式一般用多媒体呈现情境，要求受测者口头回答或表演，并进行录制(Oostrom et al., 2010, 2011)；情景面试则是主考官与受测者在面对面(或线上)的情况下问答。

封闭式是目前主流的测验形式，方便标准化处理和快速计分。然而这种形式也易受个体作答态度、猜测和应试策略的影响，受测者易从选项中获取提示，在高利害场景中存在择优作答情况，难以有效区分高胜任力个体(McDaniel et al., 2001; Robson et al., 2007)。此外，对受测者而言，选项本身含有额外的认知负荷，需阅读完所有选项并辨析含义、做比较判断，这一过程中认知能力等额外变量会对测验结果产生影响(Lievens et al., 2015; Marentette et al., 2012)。

开放式作答一定程度上可以解决这些问题，这种形式不局限于固定答案，能够给予受测者更多自由表达的空间(Finch et al., 2018)，促进受测者对主题材料的深入理解(Bacon, 2003; Rogers & Harley, 1999; Kastner & Stangla, 2011)，使其有更高的参与动机、更沉浸地做出反应(Arthur et al., 2002; Edwards & Arthur, 2007)。开放性 SJT 题项认知负荷较小，猜测被最小化，与传统多项选择式相比，具有更理想的效标关联效度(Funke & Schuler, 1998)和预测效度(Arthur, 2002; Funke & Schuler, 1998; Lievens et al., 2019)，更接近现实生活中的思考与行为过程，具有更高的生态效度和表面效度(Kjell et al., 2018)。

尽管随着技术的进步，越来越多的研究者开始探索开放式 SJT，但目前研究仍处于起步

阶段(Cucina et al., 2015)。有研究者对书面回答式(Lievens et al., 2019)和视听构建式(Oostrom et al., 2010, 2011)的测验形式进行了探索，是富有创新性的尝试，然而评分环节仍采用人工评分方式。人工评分的时间和人力成本高(Edwards & Arthur 2007; Downer et al., 2019; Iliev et al., 2015)，易受评分者效应(Rater Effects)影响(Edwards & Arthur, 2007; Lievens et al., 2019)。在 Lievens 等(2019)的研究中，评分员在每个受测者上平均花费约 35 分钟，在 Funke 和 Schuler(1998)的研究中，使用了三人评分以保证评分质量。因此，在对效率要求高的大规模施测中，这类开放式测验往往会被谨慎选用。评分问题已成为阻碍开放式 SJT 发展的重要因素(Iliev et al., 2015)，迫切需要解决自动化评分问题。

相较于人工评分，自动化评分(Automated Scoring)适用于更多元的测评任务，成本更低且能够实现即时反馈。而如何实现开放式 SJT 的自动化评分，相关研究甚少，尚未有明确的做法和系统的研究范式。Guo 等(2021)使用自然语言处理(Natural Language Processing, NLP)技术分析了五个开放式 SJT 的公开数据，采用 Doc2Vec 将文本转换为向量，使用岭回归来预测人格得分，其平均相关系数为 0.28（r=0.22~0.38)，相关性较低，也并未报告该方法的可靠性和有效性。Tavoosi (2022) 设计了包含 4 道题目的反生产工作行为(Counterproductive Work Behavior, CWB)开放式 SJT，并采用 N-gram 方法进行主题建模，抽取了主题词，但并未实现评分。

虽无明确的研究范式，但相关研究可以提供方法上的借鉴。第一，开放式 SJT 的评分标准，可以参考人工评分标准来设定。人工评分一般有简单的评分要点，再由两名以上的评分员评分，在 Lievens 等(2019)的研究中，人工评分参照了行为锚定评分表(Behavioral Anchored Rating, BAR)，评分标准更加具体、客观，该表是 Smith 和 Kendall 在 1963 年提出的，它是一种用于员工绩效评级的行为测量工具。第二，自动化评分算法。按照文本长度可以将自动化评分问题分为两类，长文本类型如作文自动化评分(Automated Essay Scoring, AES)，短文本类型如简答题自动评分(Automatic Short-answer Grading, ASAG)，开放式 SJT 自动化评分问题介于此两类之间。第三，自动评分的解释性和效度验证。心理测量学更加关注评分的可靠性、有效性和公平性，仅评分模型准确率高并不能充分说明机器评分的效果。机评效果的评估指标还包括与人工评分的相关系数、完全一致率、一致率系数(Kappa)、评分分布的一致性、相关样本评分差异 t 检验等(Ramineni et al., 2012)，Williamson 等(2012)提出机器评分的效度验证框架，包括评分结果的解释、评估、外推、概化和使用 5 个方面。

其中，上述第二点自动化评分算法是本研究的核心，以下详细介绍。AES 和 ASAG 适用的问题场景和评估重点皆不同，AES 侧重于评估文本的立意、结构、写作风格、语法和

连贯性等，开放程度高，评分核心是文本特征抽取(Rudner & Liang, 2002; Yang et al., 2022)。而 ASAG 的文本一般有若干单词或小短句，题目有参考答案，开放程度较低，是围绕标准答案的有限开放，简答题考察特定知识点，因此评分核心侧重于评估语义内容(Burrows et al., 2015)，常见方法有关键词匹配法，即作答文本的关键词越多则分数越高，或采用相似度算法，即作答文本与标准答案的相似度越高则分数越高。

不同于以上两类，开放式 SJT 的自动化评分是一个新的问题类型。主要表现为：(1)开放程度不同。既不是完全发散式(SJT 的回答可以被归类为有限的类别)，亦不存在标准答案，在相同问题情境下不同个体有着独特的解决方案，并不存在明确的、基于专业知识的"正确"答案(Whetzel & McDaniel, 2009)。(2)评分标准不同。开放式 SJT 中，文本挖掘的重点在于自然语言文本与所测心理特质的关系，某一情境下的不同做法代表着受测者不同的能力水平及特质倾向，而这种倾向的差异正是评估的重点。因而，开放式 SJT 的自动化评分很难直接参考既有算法：文本风格辨析与所测心理特质之间难以建立实际联系；关键词法更关注表层语义的相似程度，并不适用于语义更加丰富的 SJT 作答文本；相似度算法亦不适合，开放式 SJT 逻辑上并无标准答案，若采用此方法，则背离了 SJT 题目设计的初衷。

考虑到作答文本中包含不同类型的做法，可以从文本语义内容入手，尝试将评分问题转化为文本分类任务(Lubis et al., 2021; Ramesh & Sanampudi, 2022; Süzen et al., 2020)。自动文本分类(Automated Text Classification)是将文本自动划分到某些预定义类别中的过程(Basu & Murthy, 2013)。文本分类的流程主要包括：文本预处理、特征提取、模型训练、模型评估、模型优化与应用等部分。这种有监督的文本分类流程如图 1 所示，包括两个阶段，第一个阶段是在有标签的训练数据上进行模型训练，第二个阶段是应用训练好的模型对测试数据预测并作性能评估。在两个阶段中，文本数据需要进行相同的预处理和特征提取操作，例如去停用词、统计词频等，从而获取计算机可直接计算的数值型文本表征。所训练的分类模型可以看作从文本表征到分类标签的映射函数，通过指定的机器学习算法训练得到，并实现对分类文本所关联的标签做出预测。

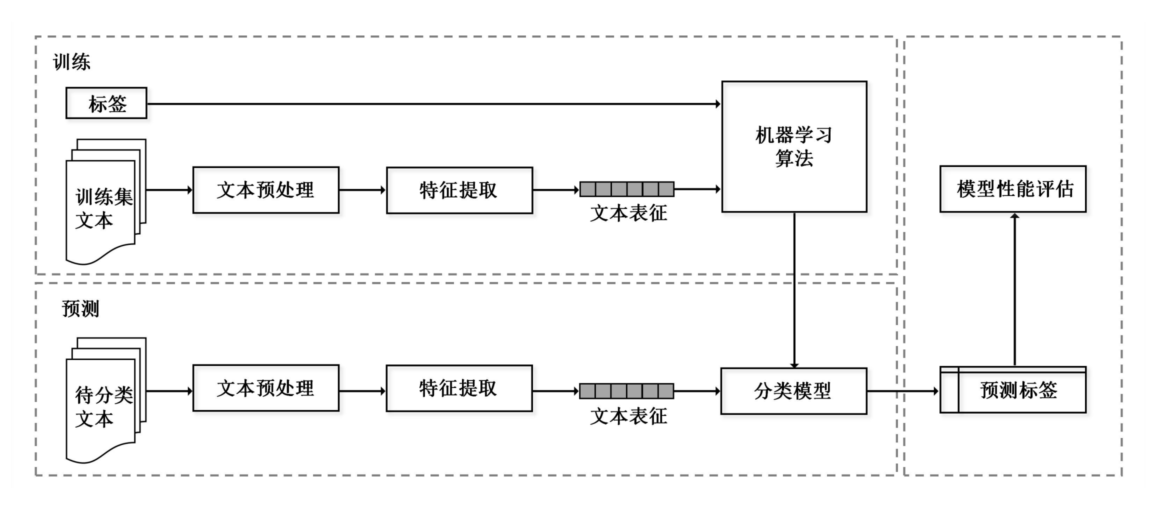


图 1 有监督的文本分类框架图

机器学习特别是深度学习模型在文本分类(Yang et al., 2022)任务中能够取得较好的结果。常用的机器学习分类算法有支持向量机(Support Vector Machines, SVM)、K 近邻(k-Nearest Neighbor, KNN)、朴素贝叶斯(Naive Bayes)、决策树(Decision Tree)等。近几年，基于深度神经网络的文本分类方法有了极大突破，展现出了更强大的性能。深度学习的方法是基于预训练的词向量模型，使用如卷积神经网络(Convolutional Neural Network, CNN)、循环神经网络(Recurrent Neural Network, RNN)等深度神经网络实现文本分类任务，在语料足够的情况下，可以表现出极佳的性能，执行文本评分任务可以达到接近人的水平，甚至比人工评分表现出更强的稳定性。

综上而言，开放式情境判断测验具有不可替代的优势，适用于需对个体进行细粒度刻画的场景中，且这类自由式作答文本中蕴含着丰富的情绪情感信息、表征着人格特质与行为偏向，对文本内容进行挖掘，可以更全面地测量个体心理，实现个性化评价。但评分问题目前存在一定困难，主要有：(1)评分标准的制定。目前评分多依赖于专家经验。(2)自动化评分的实现。评估自由文本本身就具有挑战性，在心理测评应用场景中，更是由于计算机不理解作答真实含义，使得自动化评分难以实现(Kastner & Stangla, 2011; Zhang et al., 2020)。(3)自动化评分的解释与效度验证。开放式测验的自动化评分研究较少，且难以解释评分模型输出的预测分数的含义，评分的效度验证等问题仍有待研究。

本文探索了开放式 SJT 在教师胜任力测评任务上的应用，以中小学教师为研究对象，基于心理测量学的框架开发一套开放式 SJT，结合典型行为反应设计评分标准，采用深度学习模型实现自动化评分。自动评分过程总体分为三个环节：(1)设定评分规则，在人工编码的基础上基于该情境下的关键行为，逐题确定评分规则，评分规则中包含行为反应项与对应分值；(2)自动文本分类。分别采用文档层面和句子层面的思路建模，通过实验比较多种模

型的分类效果，选用简单有效的分类模型对全部题目评分；(3)评分性能验证。从模型性能、机评信效度等多方面验证评分效果。具体流程如图 2 所示。

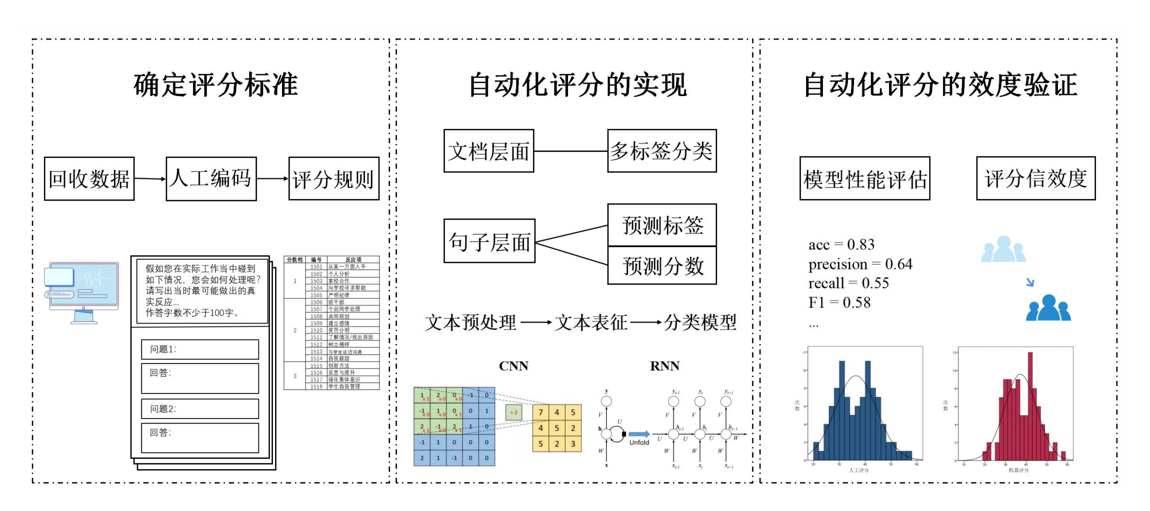


图 2 开放式情境判断测验文本自动化评分流程图

研究预期：(1)自主开发的开放式教师胜任力 SJT 信效度较好，能有效区分教师胜任水平；(2)基于深度学习的文本分类模型可应用在此类无标准答案的主观题评分任务上，机评准确性高；(3)机器评分具有较好的信效度，人机评分存在正向的强相关。

2 研究过程

2.1 被试

深圳市 627 名中小学教师参与测试(年龄：26~40 岁，$M = 31.52$ 岁，$SD = 2.2$)，其中女性 463 人，男性 164 人，语数英三科教师占 42.9%，其余学科占 57.1%。

2.2 研究工具

2.2.1 开放式教师胜任力情境判断测验

编制过程如下。

确定测验维度。采用经典流程(徐建平, 2004)构建中小学教师胜任特征，采取行为事件访谈法(Behavioral Event Interview, BEI)对北京市 8 所中小学的 12 名一线教师进行半结构化访谈，其中女性 7 人，男性 5 人，骨干教师 6 人。引导受访者回顾生涯中最成功与最遗憾的事件，每人访谈 2-3 小时。对访谈录音和文本整理后，归类汇总频次较高的关键胜任特征。最终确定胜任力模型如下，包含 4 项一级维度和 10 项二级维度：(1)学生导向：关爱学生、发展他人；(2)问题解决：动态决策、灵活应变；(3)情绪智力：理解他人、情绪控制、人际沟通；(4)成就动机：责任心、挑战困难、坚毅。

编制题目。基于文献和访谈，确定教学中的五类典型问题情境：学生管理、课堂教学、

同事相处、学生辅导、家校沟通。依据四项一级维度，选取有代表性的 54 个问题情境编制成题干与选项，统一采用指导语“在这样的情况下，你会怎么做？”。

专家评定与题目修订。向河南省 54 名教学经验丰富的小学教师发布专家评定问卷，教龄 10 年以上的占 88.24%，收回有效问卷 34 份。此版本题目为包含 4 个选项的单选题，除了完成测验，还需完成评价问卷，包括：对情境真实程度(5 点计分)做出评价；评定选项，回答实际、最优、最差、补充做法，并提出修改建议。经统计，情境真实度均值为 3.61(满分 5 分)。对选项分布进行分析，发现存在明显的优势作答倾向。根据专家意见，对试题进行修订，最终确定包含 20 道题目的开放式 SJT，分为 4 个维度：学生导向(题目号为 1、8、9、10、12、16、20)，问题解决(题目号为 3、4、6、7、17、18)，情绪智力(题目号为 2、5、11、19)，成就动机(题目号为 13、14、15)。

2.2.2 效标工具

工作满意度问卷。采用冯伯麟(1996)编制的教师工作满意量表，共 26 道题，包含自我实现、工作强度、工资收入、领导关系、同事关系 5 个维度。使用本次收集的数据作信效度检验，整体α系数为 0.89(N = 627)，五个维度的α系数分别是：自我实现 0.84、工作强度 0.76、工资收入 0.77、领导关系 0.79、同事关系 0.73。验证性因子分析结果如下：χ^2 = 1055.595，df = 289，χ^2/df = 3.65，RMSEA = 0.065，CFI = 0.868，TLI = 0.851，SRMR = 0.063。

公用教学理念与学科教学理念问卷。其中，公用教学理念问卷 12 道题，对问卷进行验证性因子分析，删去 2 个因子负荷低于 0.3 的题目(题目号为 2、12)，保留 10 道题。经分析，整体α系数是 0.88(N = 627)，模型拟合良好(χ^2 =131.363，df = 35，χ^2/df = 3.75，RMSEA = 0.066，CFI = 0.964，TLI = 0.954，SRMR = 0.029)。学科教学理念分为语文、数学、英语三科，各部分的α系数分别是 0.93(n = 99)、0.68(n = 86)、0.78(n = 84)。

综合教学水平评估材料。共 181 人提交了完整材料，由 6 名教学专家评分，每个维度满分 3 分。评估材料涵盖了教学的前中后期工作，具体包括：(1)教学设计：教师依据统一的要求提供一课时的教学设计，评价标准包括教学依据、目标、重点、难点、方法、过程 6 个方面；(2)教学视频：一个完整的 30 分钟以上的课堂教学录像，依据课堂观察量表(凌晨, 2020)，从课堂管理、教学内容、思维培养、情感关注 4 个维度对视频进行评分，量表的α系数为 0.83，四个维度的α系数依次为 0.67、0.65、0.41、0.69，验证性因子分析结果为：χ^2 = 150.12，df = 82，χ^2/df = 1.83，RMSEA = 0.075，CFI = 0.897，TLI = 0.868，SRMR = 0.060。(3)学生作业：布置作业并按照优良差各 3 份提交共 9 份具有代表性的学生作业；由教学专家对教师的作业内容设计、作业评价标准设计以及对学生作业的分析 3 个部分进行评分。

2.3 数据分析

使用 SPSS 26.0 和 Mplus 8.3 做测验质量分析，使用 Nvivo 11 软件进行人工编码，使用 Python 3.8 进行数据训练和预测。

2.4 确立评分标准

2.4.1 问题界定

文本评分首先要考虑的是评分标准问题。开放式 SJT 作答文本的特点为：一个问题的回答中包含若干种做法，且含解决问题的步骤、逻辑与顺序等，不具有单一的、明确的答案。评分的核心不是该情境下的做法是否正确，而是文本中的典型行为反应模式与教师胜任力模型的契合程度。在本研究中，不设定答案模板，依据行为锚定(Behavioral Anchored)评分思路，关注情境中的特定刺激引发的关键行为，由编码员为文本中的所有回答分类，并将类别进一步聚类为典型行为反应集，并为反应项赋予分值。由于不同场景下的关键行为不同，每道题需单独设定评分规则。

2.4.2 人工编码

选取 300 人的作答文本进行人工编码，剔除作答时间少于 1000 秒与明显不认真作答(重复或无关文字)的 10 人，保留 290 份文本。选用 4 名心理学专业的研究生编码，编码前统一接受半天培训，培训内容包括测评维度、编码标准、软件操作、遇争议项的处理原则等，题目被随机分配。

编码流程包括两部分：第一，确定行为反应项。具体地，每道题由两名编码员先通读文本，独立梳理被试所有的行为反应项，再一同修改合并，为反应项聚类，以确立典型行为反应项(10~30 类，多为十几类)。第二，人工编码标注(打标签)。一名编码员在 Nvivo 软件中逐句标注，另一名编码员对编码结果核查，过程中可以提出不同意见，也可继续对编码规则合并完善；每道题编码完成后导出结果，按照 ID 整理句子标注数据集。

2.4.3 制定评分规则

由上一步骤得到评分规则中的行为反应项，接下来，为反应项赋予分值。依据作答结果与胜任力特征的匹配程度，为更贴近胜任特征的反应项赋予更高分值，同时关注行为的丰富度、具体性、全面性、逻辑性等所体现的思维水平和能力的差异，对各个反应项赋分，分值即为权重分数，权重分数需能够有效地体现行为反应的差异。采用 3 分制(0~3 分)，1=差，3=优秀，0=偏题或无效作答。每道题的赋分皆由两名评分员讨论后确定，直至达成一致。

2.4.4 分数合成

基于人工编码环节得到的每个 ID 的行为反应项，依据制定评分规则环节得到的反应项

对应的分值，逐题将每个 ID 的各行为反应项转为分数。一段作答中一般包含多个行为反应项，将其分数加和后合成单道题原始总分，根据百分位数换算成等级分数，前 27%等级分数为 3 分，后 27%为 1 分，得到每道题的得分。除此之外，还计算了维度分数和测验总分(即 20 道题的分数加和)。题目采取 3 分制(1~3 分)，满分 60 分。

2.5 自动化评分的实现

2.5.1 数据集和评价指标

选定 ID 为 1~300 的已标注文本作为数据集，共 20 道题目，6000 道回答。每道题的文本中，按照 300 人的 2:1 划分训练集和测试集。在机器学习领域，对分类任务的评价一般采用准确率(Accuracy, Acc)、精确率(Precision, P)、召回率(Recall, R)、F1 值等评价指标。下面以二分类为例，对四个指标的计算过程进行说明。假设二分类包括正类和负类，表 1 为二分类情况下的混淆矩阵，矩阵中的元素定义为：1)TP(True Positive)：实际为正类且预测为正类的样本个数；2)TN(True Negative)：实际为负类且预测为负类的样本个数；3)FP(False Positive)：实际为负类且预测为正类的样本个数；4)FN(False Negative)：实际为正类且预测为负类的样本个数。

表 1 二分类的混淆矩阵表

	预测正例	预测反例
实际正例	TP 真正例	FN 假负例
实际反例	FP 假正例	TN 真负例

准确率反映模型在所有样本上的预测性能，等于分类正确的样本数除以总体样本数，即混淆矩阵中的对角线元素之和除以矩阵中所有元素之和，即准确率 Acc=(TP+TN)/(TP+FN+FP+TN)。精确率、召回率和 F1 值三个指标在每个类别上需单独计算。以二分类中的正类为例，精确率等于将正类样本预测为正类的数量除以所有预测为正类的样本数量，即 P=TP/(TP+FP)；召回率等于将正类样本预测为正类的数量除以真实的正类样本数量，即 R=TP/(TP+FN)；F1 值为精确率和召回率的调和平均值，即 F1=2PR/(P+R)。本文中的文本分类主要为多分类任务，在计算评价指标时先分别在每个类别上计算 P、R、F1，然后根据每个类别的样本数量计算加权平均值得到最终的精确率、召回率和 F1 值的评估结果。

2.5.2 文档层面多标签文本分类

传统的文本分类任务多是单标记学习，每个文本只隶属于一个类别标签，在一个类别上标记互斥，用 0 或 1 来标记，但实际许多样本同时属于多个类别的多个标签。Schapire 于 1999

年提出了多标记学习，从标签集合中为每个实例分配最相关的类标签子集。根据数据集一段作答文本中同时包含多类行为反应项的特点，首先基于文档层面尝试多标签(Multi-label)分类方法。

选用第一题作为实验，使用深度学习算法进行分类建模，从而实现从作答文本到标签体系的自动化映射。在具体操作中，先进行数据预处理，去除停用词，输入文本，通过 Jieba 分词和 Word2vec 预训练词向量转化为数字矩阵形式，再连接具有可训练参数的神经网络层、全连接层和 SoftMax 层，最终输出文本所属各个标签的概率。其中，在神经网络层应用了多种深度学习方法，包括卷积神经网络(Convolution Neural Network, CNN)(Kim, 2014)、循环神经网络(Recurrent Neural Network, RNN)(Zhao et al., 2019)、循环神经网络串联卷积神经网络(Recurrent Convolution Neural Network, R-CNN)(Lai et al., 2015)和循环神经网络串联注意力网络(RNN + Attention)(Pang et al., 2021)。其中，CNN 主要通过卷积核参数来捕捉各类标签的文本局部深度特征；RNN 通过循环单元结构来捕捉各类标签的文本全局深度特征；R-CNN 同时发挥两者的优势将 RNN 和 CNN 进行串联使用；Attention 则通过神经网络计算文本中每个词的权重来优化文本深度表征，通常与 RNN 进行串联使用。

2.5.3 句子层面文本多分类

多标签分类任务是在文档层面对整段作答文本直接输出多个标签，如果将文档拆分，可以在句子层面输出每句话单独的标签。在人工编码环节，已得到逐句编码的标注集。

随机选取四道题目，首先进行数据预处理，通过“。.！？；;”等标点符号和“一（一）1(1)①”等序号来分割句子，去除停用词，通过 Jieba 分词和 Word2vec 预训练词向量将文本转化为数字矩阵，使用卷积神经网络(CNN)、长短时记忆网络(Long Short-Term Memory Neural Network, LSTM)(Hochreiter et al.,1997)、卷积神经网络串联长短时记忆网络(C-LSTM)、长短时记忆网络串联注意力网络(LSTM+attention)四种深度学习模型进行训练，分别做分数预测和行为反应项预测。其中 LSTM 能够有效应对梯度消失、梯度爆炸问题，它是 RNN 的结构变种；C-LSTM 是 CNN 与 LSTM 的结合(Zhou al.,2015)，既能获得句子的局部特征，也可以获取全文中的时态句子语义。模型通过学习标注集中每种反应项或分数对应的句子集合，找到文本之间的深层语义关系，以此完成模型训练。每个句子皆输出两种预测结果，一是分数预测，即输出句子的分值(0~3 分)，二是标签预测(行为反应项)，可以帮助更细致地评估作答者的思想和能力。

3 结果

3.1 评分规则

原始作答数据集中，每道题作答文本 100~300 字，20 道题共 1353365 字。取前 300 份进行编码，已编码 647322 字，单题标注 724~1453 句，总计标注 19368 个句子。选取第一道题做编码一致性检验，两个评分者的人工编码一致性 $r = 0.84$，二次加权 Kappa 系数为 0.78。每道题的评分规则在人工编码后产生，主要包含两大部分——此情境下的典型行为反应项以及分值，每个反应项有唯一的编号，共形成 20 个评分规则。

3.2 测验质量分析

以多种信度指标来考察多维测验信度(顾红磊, 温忠麟, 2017)。经计算，在双因子结构下，即把胜任力作为全局因子，四个维度作为四个局部因子，同质性系数(Homogeneity Coefficient, HC)和总合成信度分别为 0.88 和 0.96。测验整体的 Cronbach's α系数为 0.91，各维度的α系数为：学生导向 0.79，问题解决 0.76，情绪智力 0.66，成就动机 0.60。

为检验测验的结构效度，设定并比较了四种验证性因子分析模型：M_1 为单因子模型，即所有题目负载于一个因子；M_2 为四因子模型；M_3 为双因子模型(Bi-factor Model, BFM)，即在 M_2 基础上，所有题目还负载于一个全局因子，全局因子与局部因子互不相关；M_4 为双因子模型，全局因子与局部因子不相关，局部因子两两相关。结果见表 2，M_4 明显优于其他模型，因此选定 M_4 为最佳模型，测验具有较清晰的双因子结构，具有一个胜任力全局因子和四个维度。

表 2 教师胜任力情境判断测验的验证性因子分析(n = 290)

模型	χ^2	df	χ^2/df	CFI	TLI	SRMR	RMSEA
M_1	264.34	170	1.56	0.947	0.941	0.042	0.043
M_2	256.12	164	1.56	0.948	0.940	0.041	0.044
M_3	226.59	190	1.19	0.957	0.946	0.038	0.042
M_4	179.58	144	1.25	0.980	0.974	0.033	0.029

采用工作满意度、教学理念、教学能力作为效标来检验效标关联效度，结果见表 3，胜任力总分与工作满意度(r_1=0.20，p = 0.001)、公用教学理念(r_2=0.21，p＜0.001)、学科教学理念(r_3=0.22，p＜0.001)、教学能力中的教学设计(r_4=0.26，p＜0.001)、课堂评价(r_5=0.20，p = 0.007)、学生作业(r_6=0.22，p = 0.003)皆呈显著相关。

表 3 教师胜任力总分及其维度与效标变量的相关分析表

变量	*M*±*SD*	1	2	3	4	5	6	7	8
n = 290									
1 总分	37.77±8.17	1							
2 学生导向	1.89±0.46	0.89***	1						
3 问题解决	1.87±0.48	0.87***	0.68***	1					
4 情绪智力	1.91±0.50	0.78***	0.58***	0.55***	1				
5 成就动机	1.89±0.54	0.75***	0.55***	0.58***	0.54***	1			
6 工作满意度	3.88±0.29	0.20**	0.22***	0.14*	0.15*	0.13*	1		
7 公用教学理念	4.58±0.41	0.21***	0.18**	0.15*	0.20***	0.18**	0.45***	1	
8 学科教学理念	3.58±0.49	0.22***	0.21**	0.18**	0.13*	0.20**	0.33***	0.50***	1
n = 181									
1 总分	38.95±8.30	1							
2 学生导向	1.92±0.49	0.87***	1						
3 问题解决	1.96±0.47	0.88***	0.68***	1					
4 情绪智力	1.95±0.52	0.76***	0.53***	0.54***	1				
5 成就动机	1.97±0.61	0.71***	0.52***	0.55***	0.49***	1			
6 教学设计	2.60±0.31	0.26***	0.24**	0.27***	0.16*	0.18*	1		
7 课堂视频	2.54±0.32	0.20**	0.21**	0.17*	0.10	0.13	0.61***	1	
8 学生作业	2.52±0.44	0.22**	0.18*	0.22**	0.18*	0.16*	0.43***	0.39***	1

注：$^{*}p<0.05$，$^{**}p<0.01$，$^{***}p<0.001$；教师胜任力测验单道题采取 3 分值，测验总分满分 60 分；工作满意度、教学理念问卷采取 5 分值；教学设计、课堂视频、学生作业的专家评分各维度采取 3 分值。

3.3 自动化评分模型性能

3.3.1 文档层面文本多分类

使用多标签标记方法，将整段回答输出多标签结果。实验结果如表 4 所示，各模型在测试集上的表现皆不够理想，准确率为 46%~55%。研究者推测，一方面是受限于样本数量，另一方面，是由于分类类别过多，题目类别平均包含 20 类左右，且大多数标签是仅有少数标注数量的尾标签。

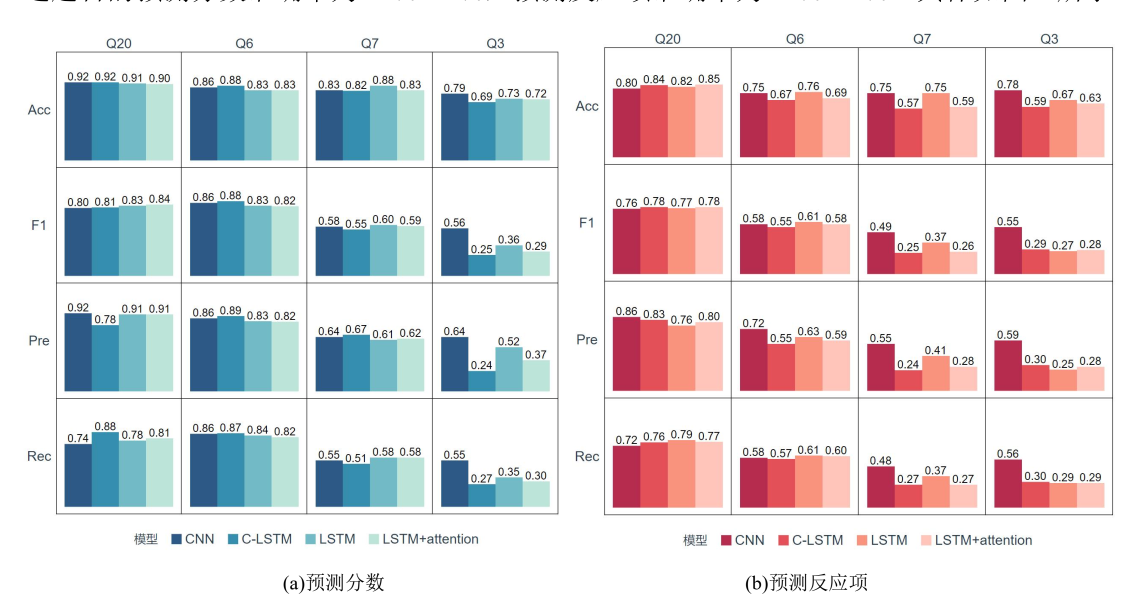
表 4 文档层面多标签文本分类的模型实验结果对比表

模型	Accuracy	Precision	Recall	F1
CNN	0.46	0.51	0.25	0.58
RNN	0.51	0.56	0.36	0.59
R-CNN	0.55	0.71	0.36	0.67
RNN+Attention	0.48	0.60	0.35	0.60

3.3.2 句子层面文本多分类

将作答文本拆分为句子单元，在随机选取的四道题目上进行模型训练。实验结果表明：(1)对于分数预测任务，在题 20 上，四种算法的准确率、F1 值差距较小，CNN 的精确率最高，C-LSTM 的召回率最高；在题 6 和题 7 上，四种算法的四个指标差异较小，题 6 中 C-LSTM 略好，题 7 中 LSTM 略好；在题 3 上，CNN 明显优于其他模型。(2)对于反应项预测任务，在题 20 上，四种算法的准确率、F1 值差距较小，CNN 的精确率较高，LSTM 的召回率较

高；在题 6 上，四种算法的 F1 值、召回率差距较小，CNN 和 LSTM 的准确率较高，CNN 的精确率最高；在题 7、题 3 上，CNN 的四项指标皆最佳。综合来看，CNN 表现最好，四道题目的预测分数准确率为 79%~92%，预测反应项准确率为 75%~80%。具体如图 3 所示。



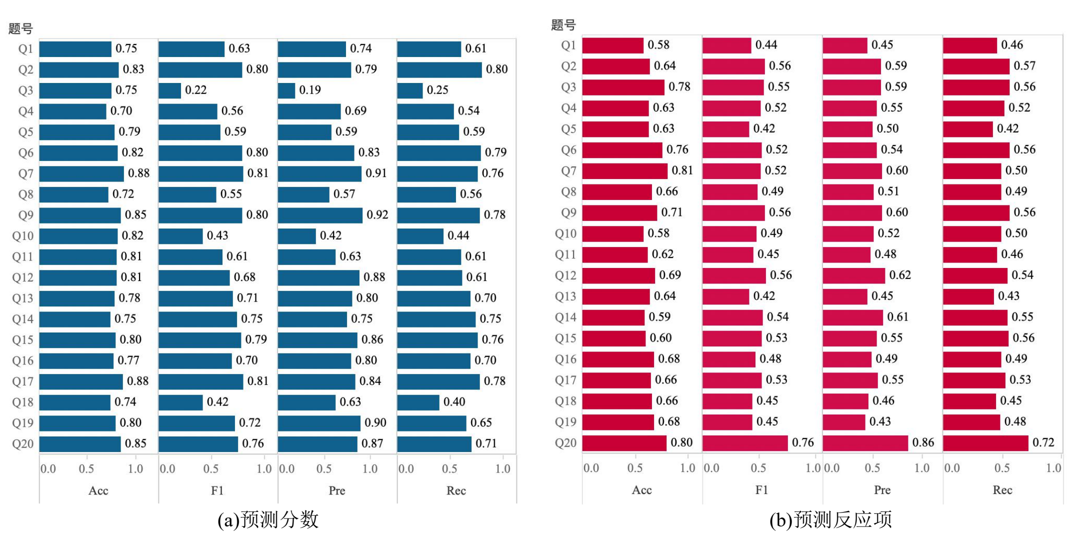
(a)预测分数　　　　(b)预测反应项

图 3 四种模型在四道题目上预测反应项和预测分数任务的结果对比图

注：Acc 为准确率(Accuracy); F1 为 F1-score; Pre 为精确率(Precision); Rec 为召回率(Recall)，下同。

3.3.3 整体性能

句子层面的准确率高于文档层面，因此采用句子层面文本多分类的方法，选定综合表现最佳的 CNN 模型对所有题目进行自动评分。结果如图 4 所示，计算机在 20 道题上预测分数的准确率为 70%~88%，结果较好；预测行为反应项的准确率为 58%~81%，考虑到数据集训练语料量较少而语义又具有丰富性的特点，以及分类类别较多，为十几至二十几类，故此准确率仍属较不错的结果。



(a)预测分数　　　　(b)预测反应项

图 4 CNN 在 20 道题上的结果图

3.4 自动化评分的效度验证

将标注集前 200 人的数据作为训练集，后 100 人的数据作为测试集。100 人的机器评分结果中，删去数据不完整及作答时间过短的 6 人，对 94 人的 1880 道作答的人机评分结果进行对比分析，检验机器评分的信效度。

3.4.1 人机评分一致性

人机评分的数据分布。人工评分与机器评分的总体数据分布形态接近，人工评分总分(36.36±7.99)的峰度为-0.592，偏度为 0.175，机器评分总分(37.23±7.83)的峰度为-0.345，偏度为 0.151，如图 5 所示。

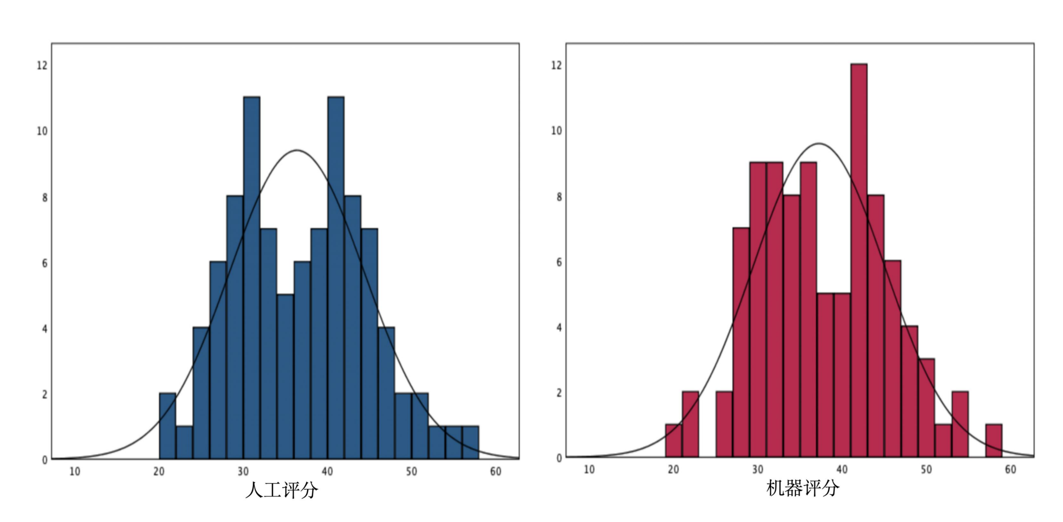


图 5 人工评分与机器评分总分的分数频率分布图

相关性。采用相关系数作为评价标准，有研究者指出机器评分与人工评分的相关系数至少达到 0.7 才可用于大规模、高权重考试(Ramineni et al., 2012)。经计算，人工评分总分(36.36±7.99)与机器评分总分(37.23±7.83)呈高度正相关($r = 0.95$，$p < 0.001$)，且在学生导向、问题解决、情绪智力、成就动机这 4 个维度人评(1.81±0.45、1.82±0.49、1.87±0.49、1.78±0.54)与机评(1.89±0.44、1.84±0.47、1.87±0.47、1.82±0.53)皆为高度正相关($r_{学生导向} = 0.91$，$r_{问题解决} = 0.90$，$r_{情绪智力} = 0.81$，$r_{成就动机} = 0.89$，$p < 0.001$)，达到大规模考试的使用要求。在 20 道题目上，人机评分的相关系数 r 依次为 0.88、0.64、0.80、0.71、0.78、0.60、0.88、0.63、0.48、0.82、0.84、0.54、0.84、0.81、0.85、0.74、0.68、0.75、0.65、0.90，$p < 0.001$。

一致率系数(Kappa)。采用二次加权 Kappa 系数(QWK)作为评价标准，Williamson 等(2012)认为自动评分的 QWK 应至少为 0.7 才能用于高风险测试情况。本研究中人工评分与机器评分的 QWK——总分(0.82)和各维度(学生导向 0.89、问题解决 0.90、情绪智力 0.81、成就动机 0.89)皆已达到用于高风险测验的标准。

3.4.2 机器评分的信度与效度

采用 Cronbach's α系数来衡量测验内部一致性信度，使用机评结果计算得出，测验整体α系数为 0.87，各维度为：学生导向 0.66，问题解决 0.73，情绪智力 0.55，成就动机 0.55。验证性因子分析(n = 94)结果显示，χ^2 = 210.896，df =164，χ^2/df = 3.75，RMSEA=0.055，CFI=0.884，TLI=0.866，SRMR=0.029，各项目在各因子上的因子载荷在 0.412~0.659 之间，结构效度不如人工评分。机器评分的总分与各维度的相关系数如表 5 所示，各维度与总分存在较高相关，各维度间存在中等水平相关。效标关联效度如表 5 所示，胜任力总分与公用教学理念呈显著相关(r = 0.22，p = 0.036)。

表 5 机器评分的描述性统计和相关分析表(n=94)

变量	$M±SD$	1	2	3	4	5	6	7	8
1 机评总分	37.23±7.83	1							
2 机评学生导向	1.89±0.44	0.90***	1						
3 机评问题解决	1.84±0.47	0.88***	0.69***	1					
4 机评情绪智力	1.87±0.47	0.80***	0.63***	0.58***	1				
5 机评成就动机	1.82±0.53	0.72***	0.53***	0.54***	0.51***	1			
6 公用教学理念	3.85±0.27	0.22*	0.11	0.22*	0.24*	0.19	1		
7 学科教学理念	4.28±0.39	0.21	0.17	0.27*	0.19	0.01	0.63***	1	
8 工作满意度	3.65±0.58	0.13	0.15	0.09	0.10	0.08	0.32**	0.44***	1

4 讨论

本研究试图探索一种以心理测量学理论为基础的开放式情境判断测验自动化评分范式，为了验证自动评分的有效性，聚焦具体的研究问题——教师胜任力测评，开发了开放式 SJT，设置细粒度的评分规则，使用 NLP 技术进行文本特征识别和分类，分别在文档和句子层面使用 CNN、RNN、R-CNN、RNN+Attention、LSTM、C-LSTM、LSTM +Attention 等多种深度学习模型对开放式 SJT 的自动化评分方法进行了探索。结果显示，句子层面的分类效果优于文档层面，其中，CNN 表现较好，模型预测分数的准确率达到了 70%~88%，预测反应项的准确率为 58%~81%，模型性能较好，能够对文本进行较准确的自动评分，下面进行具体讨论。

4.1 评分标准的设计

确立评分标准前需进行问题界定，根据测验的内容和类型、作答文本的特点来确定评分策略。比如有无标准答案决定着评分逻辑和算法设计，作答文本的长度与语义丰富度决定着是否需要人工编码的参与，也决定着计分策略。评分规则应尽可能体现作答者个人特质层面的信息，重点考虑两个问题：(1)合理分类。行为反应项全面、具体、有代表性，尽可能涵

盖所有类型。类别既需充分体现差异，又要避免分类过细带来的随机性。反应项若细致具体，则更能体现差异和区分性，但由于类别过多，评分的准确率会降低；若行为反应项少，则有利于提高预测准确率，但会导致区分度降低。(2)合理赋分。通过分数高低体现作答者水平高低，各行为反应项的赋分是较困难的过程，需反复斟酌对比，综合考量。

另一方面，测验质量对评分效果有着直接的影响。测验开发与自动评分这两个环节并不是独立的，本文探索的自动评分方法，关键不在于分类模型的复杂或先进，并不着重于追求更完美的模型，而在于设计一套可行的开放式 SJT 自动评分方法，设定合理的评分规则、选取合适的评分模型，在此基础上逐步提高模型评分的准确率。开放作答并不意味着测验开发过程的随意和自由，不可随意设置题目，而是应该依据规范的测验开发标准，在一套合格的、信效度较好的测验基础上实现评分自动化。测验开发者需对测评维度有深刻理解，在大量的调研、访谈中把握所测特质的内涵和行为表现，在此基础上才能设计良好的评分规则。此外，题目需注意用词，避免出现歧义、过多的额外或干扰信息，影响测验质量。

4.2 自动化评分过程

研究中使用了多种方法、选用多种模型进行实验对比，以选择最优模型。根据具体任务的输入输出形式，自动化评分有多种建模思路，在实践中一般需进行多种尝试并选取更简单有效的建模方法。本研究中，输入为被试的作答文本，输出为该文本涉及的多个反应项或多级评分，这种输入输出形式可以直接对应机器学习领域的多标签分类任务，因此首先尝试了文档层面的多标签文本分类，这种建模方法没有引入句子级别的标注信息，如果能够达到可用的性能可优先使用。然而，在实践中，多标签分类结果欠佳，文章仅以第一题为例说明了这个过程，而句子层面的自动化评分能够取得更有效的结果，因而采用了这种思路。

不同类型的深度学习模型在处理文本分类任务时，具有独特的优势和限制，这些特点会在自动化评分性能方面产生不同的影响。例如，卷积神经网络(CNN)在文本中主要捕获局部特征，如词组、短语等，对于需要考虑长程依赖关系的任务，CNN 可能表现较差，因为它无法有效地处理长文本序列中的全局信息；循环神经网络(RNN)及其变体，如长短时记忆网络(LSTM)和门控循环单元(GRU)，在处理序列数据时能够捕捉上下文信息，适用于对文本中长期依赖关系较强的任务。然而，传统的 RNN 难以处理长序列，虽然 LSTM 和 GRU 在一定程度上缓解了这些问题，但仍受到文本长度的限制而在一些文本分析任务上表现较差；注意力机制(Attention)使模型能够在处理文本时聚焦于关键部分，有助于更好地捕捉重要信息，但早期的注意力机制通常与 RNN 绑定使用，容易受到 RNN 模型的限制。本研究中，由于句子层面的反应项分类任务通常与特定的词组、短语等相关联，因此 CNN 在本研究上

的性能最优是可以理解的。在广泛的研究任务中，不同类型的深度学习模型在自动化评分中具有各自的特点，模型的选择将对评分性能产生显著影响。根据任务的特殊要求，结合模型的优势和限制，选择合适的模型有助于提高自动化评分的准确性。此外，在预训练语言模型(Pre-trained Language Model)、大语言模型(如 ChatGPT)等被提出后，自动化评分模型也有了更丰富的选择，但考虑到场景的特异性，仍然需要经过严谨的性能评估、信效度检验才能确定评分模型的可用性。

自动评分的效果亦受多种人工因素影响。在评分前需做好数据预处理，注意分句方式。句子的分割效果直接影响着分数，对于机器来说，区分一段文本中的不同语义单元较为困难，数据集中标点符号使用不规范的现象越多，分句质量越差。因此，在机器分句之后增加一个对分割数据集的校验工作，会有助于后续获得更好的评分效果。在更广泛的测验类型中，应根据文本长短、语义复杂度，选用合适的分句标志或符号。此外，也应采用多种方式保障人工编码的质量，优化评分规则的设置。

4.3 自动化评分的效度和可解释性

人工评分和机器评分在心理测验的使用中表现出了不同的特点。王妍和彭恒利(2019)对比了人机评分的特点，在辨别考生作答偏题、背诵模板、对考生的作答进行语意判断、识别作答语序和逻辑顺序这几个方面人类表现得较好，而机器评分具有更少的评分趋中现象，对考生作答的整体把握能力和识别异型卷的能力都更强。机器评分的结果能否辅助甚至代替人工，除了关注预测准确率等一系列模型评价指标，也需关注评分信效度，尤其要做评分的效度验证。本研究中，在测验总分和四个维度分上，人机评分的 r 和 QWK 皆大于 0.8，在部分题目上机器评分比人工评分具有更强的稳定性，如第一题人机评分一致性高于两个评分员之间的一致性($r_{人\text{-}机}= 0.88$, $r_{人\text{-}人}= 0.78$)。因此，自动化评分系统是有效的，在阅卷过程中可以至少替代一位评分员进行评分，实现人机结合评分或自动化评分。

效度研究被视为一种对测验分数作出可接受的(Plausible)解释的过程(谢小庆, 2013)，自动化评分的可解释性问题更是研究的难点。机器学习的过程通常建立一个可解释性不强的黑盒模型，难以满足一个心理测验对测评要素的描述需求，仅从数据本身、文本表征的距离远近来实现机器评分是不够的。本研究中，构建评分模型时引入了专家知识，这也是将机评过程转换成“白盒”模型的关键，将不可见的评分过程，转化为先将文本分类到评分规则里对应的行为反应项上，不仅能得到分数，也能得到作答者的行为反应项，基于这些行为项，还可对作答者的行为模式、思维方式、人格特点等作进一步挖掘。这种关注到具体行为的评分能够更细致地刻画被试的行为差异，是更细粒度的评分模型，也更具有可解释性。

4.4 实践启示

研究具有广阔的应用前景和实际意义，主要体现在以下几个方面：第一，对 SJT 的开放式作答形式进行探索，减少选择题形式 SJT 的被试猜测和作假行为，实现对个体更个性化的评估。第二，是对中文无标准答案主观题自动化评分技术的探索，搭建了开放式测验开发——分类归集——标定编码——专家赋分——自动评分——效果验证这样一个完整的范式。不再局限于简单语义计算和相似度计算，而更注重文本与所测心理特质的对应关系，通过人工编码和细粒度的评分规则，增强了自动化评分过程的可解释性。第三，自动化评分模型的准确率高、效果较好，在实践中，评分更高效、节省人力和时间，是准确、可靠的评估工具。第四，拓展开放式 SJT 的应用，也为其他类型开放式题目的自动化评分提供参考和指导，有助于将开放式题型应用在更广泛的测验场景中。

4.5 研究局限和展望

研究也存在着一些局限。具体表现在：第一，样本代表性，被试选取的是深圳市中小学各科教师，属于教育资源优质地区的青年教师，同质性较强，不能代表更一般化的教师群体。第二，标注数量有限，对于机器评分来说，评分精度受限于标注样本数量，由于时间及人力等限制条件，每题的句子标注数量在 1000 句左右，且有标签不均匀的情况，影响机器学习的效果。第三，评分规则中的行为反应项还可以尝试进一步的分类概括和调整。第四，效标选取，应选取更贴合的效标，多方面、多证据验证评分效度。

未来研究中，将考虑继续扩充题目库，不断更新时代背景下教学中出现的新问题情境，同时增强题目的针对性。同时，使用 AI 自动编码的方式来辅助归类反应项，提高效率。在评分算法上尝试更多小样本学习的方法，进一步提高机评准确率。此外，对于更丰富的开放式构答形式如语音、肢体动作等，可参考 AI 面试系统(Lee & Kim, 2021)的技术思路，探索开放式 SJT 更广阔的应用空间。

5 结论

在本研究条件下，主要得出如下结论：(1)开放式情境判断测验可从关键行为反应项上设定评分规则，自动化评分的步骤包括：行为反应项分类归集——标定编码——专家赋分——自动化评分——效果验证；(2)评分算法可从文档层面和句子层面分别设计，本研究中句子层面的文本分类效果优于文档层面，其中卷积神经网络的分类准确率较高，能对关键行为反应项的字词特征进行更好的捕捉；(3)所开发的评分模型具有稳定的效果，机器评分与人工评分一致性高，具有较好的信效度，在实践中可部分代替人工完成评分任务。

参考文献

Arthur, W., Edwards, B. D., & Barrett, G. V. (2002). Multiple-choice and constructed response tests of ability: Race-based subgroup performance differences on alternative paper-and-pencil test formats. *Personnel Psychology*, *55*(4), 985–1008.

Bacon, D. R. (2003). Assessing learning outcomes: A comparison of multiple-choice and short-answer questions in a marketing context. *Jounal of Marketing Education*, *25*(1), 31–36.

Basu, T., & Murthy, C. A. (2013, December). Effective text classification by a supervised feature selection approach. *IEEE 12th International Conference on Data Mining Workshops(ICDM),* 918–925, Brussels, Belgium.

Burrus, J., Betancourt, A., Holtzman, S., Minsky, J., MacCann, C., & Roberts, R. D. (2012). Emotional intelligence relates to well-being: Evidence from the situational judgment test of emotional management. *Applied Psychology: Health and Well-Being*, *4*(2), 151–166.

Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *Int J Artif Intell Educ, 25*, 60–117.

Cucina, J. M., Su, C., Busciglio, H. H., Thomas, P. H., & Peyton, S. T. (2015). Video-based testing: A high-fidelity job simulation that demonstrates reliability, validity, and utility. *International Journal of Selection and Assessment, 23*(3)*,* 197–209.

Downer, K., Wells, C., & Crichton, C. (2019). All work and no play: A text analysis. *International Journal of Market Research*, *61*(3), 236–251.

Edwards, B. D., & Arthur, W., Jr. (2007). An examination of factors contributing to a reduction in subgroup differences on a constructed-response paper-and-pencil test of scholastic achievement. *Journal of Applied Psychology*, *92*(3), 794–801.

Finch, W. H., Finch, ME. H., Mcintosh, C. E., & Braun, C. (2018). The use of topic modeling with latent dirichlet analysis with open-ended survey items. *Translational Issues in Psychological Science*, *4*(4), 403–424.

Funke, U., & Schuler, H. (1998). Validity of stimulus and response components in a video test of social competence. *International Journal of Selection and Assessment*, *6*(2), 115–123.

Gu, H. L., & Wen, Z. L. (2017). Reporting and interpreting multidimensional test scores: A bi-factor perspective. *Psychological Development and Education*, *33*(4), 504–512.

[顾红磊, 温忠麟. (2017). 多维测验分数的报告与解释：基于双因子模型的视角. *心理发展与教育*, *33*(4), 504–512.]

Guo, F., Gallagher, C. M., Sun, T., Tavoosi, S., & Min, H. (2021). Smarter people analytics with organizational text data: Demonstrations using classic and advanced NLP models. *Human Resource Management Journal*. Advance Online Publication.

Iliev, R., Deghani, M., & Sagi, E. (2015). Automated text analysis in psychology: methods, applications, and future developments. *Language and Cognition*, *7*(2), 265–290.

Kastner, M., & Stangla, B. (2011). Multiple choice and constructed response tests: Do test format and scoring matter? *Procedia-Social and Behavioral Sciences. 12*, 263–273.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *Proceedings of the 19th Conference on Empirical Methods in Natural Language Processing,* 1746–1751.

Kjell, O. E., Kjell, K., Garcia, D., & Sikstrom, S. (2018). Semantic measures: using natural language processing to measure, differentiate, and describe psychological constructs. *Psychological Methods*, *24*(1), 92–115.

Lai, S., Xu, L., Liu, K., & Zhao, J. (2015). Recurrent convolutional neural networks for text classification. *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 2267–2273.

Lee, B. C., & Kim B.Y. (2021). Development of an AI-based interview system for remote hiring. *International*

Journal of Advanced Research in Engineering and Technology, *12*(3), 654–663.

Lievens, F., De Corte, W., & Westerveld, L. (2015). Understanding the building blocks of selection procedures: effects of response fidelity on performance and validity. *Journal of Management*, *41*(6), 1604–1627.

Lievens, F., Sackett, P. R., Dahlke, J. A., Oostrom, J. K., & De Soete, B. (2019). Constructed response formats and their effects on minority-majority differences and validity. *Journal of Applied Psychology*, *104*(5), 715–726.

Ling, C. (2020). *Development of Classroom Observation Scale to Promote the Professional Development of New Teachers* (Unpublished master's thesis). Beijing Normal University.

[凌晨. (2020). *课堂观察量表的开发——促进初任教师专业发展* (硕士学位论文). 北京师范大学.]

Lubis, F. F., Mutaqin, Putri, A., Waskita, D., Sulistyaningtyas, T., Arman, A. A., & Rosmansyah, Y. (2021). Automated short-answer grading using semantic similarity based on word embedding. *International Journal of Technology*. *12*(3), 571–581.

Marentette, B. J., Meyers, L. S., Hurtz, G. M., & Kuang, D. C. (2012). Order effects on situational judgment test items: A case of construct-irrelevant difficulty. *International Journal of Selection and Assessment, 20*(3), 319–332.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. (2007). Situational judgment tests, response instructions, and validity: a meta-analysis. *Personnel Psychology, 60*(1), 63–91.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology*, *86*(4), 730–740.

McDaniel, M. A., Psotka, J., Legree, P. J., Yost, A. P., & Weekley, J. A. (2011). Toward an understanding of situational judgment item validity and group differences. *Journal of Applied Psychology*, *96*(2), 327–336.

Oostrom, J. K., Born, M. P., Serlie, A. W., & Van der Molen, H. T.(2010). Webcam testing: Validation of an innovative open-ended multimedia test. *European Journal of Work and Organizational Psychology*, *19*(5), 532–550.

Oostrom, J. K., Born, M. P., Serlie, A. W., & Van der Molen, H. T. (2011). A multimedia situational test with a constructed-response format: Its relationship with personality, cognitive ability, job experience, and academic performance. *Journal of Personnel Psychology, 10*(2), 78–88.

Oostrom, J. K., Born, M. P., Serlie, A. W., & van der Molen, H. T. (2012). Implicit trait policies in multimedia situational judgment tests for leadership skills: Can they predict leadership behavior? *Human Performance*, *25*(4), 335–353.

Pang, N., Zhao, X., Wang, W., Xiao, W., & Guo, D. (2021). Few-shot text classification by leveraging bi-directional attention and cross-class knowledge. *Science China Information Sciences. 64*(3).

Qi, S. Q., & Dai, H. Q. (2003). The Property Function and the Development of Situational Judgment Tests. *Psychological Exploration*, *23*(4), 42–46.

[漆书青, 戴海琦. (2003). 情景判断测验的性质、功能与开发编制. *心理学探新*, *23*(4), 42–46.]

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, *55*(3), 2495–2527.

Ramineni, C., Trapani, C. S., Williamson, D. M., David, T., & Bridgeman, B. (2012). *Evaluation of the e-rater® scoring engine for the GRE® Issue and Argument Prompts* [EB/OL].

Robson, S. M., Jones, A., & Abraham, J. (2007). Personality, faking, and convergent validity: a warning concerning warning statements. *Human Performance*, *21*(1), 89–106.

Rogers, W. T., & Harley, D. (1999). An empirical comparison of three-and four-choice items and tests: Susceptibility to testwiseness and internal consistency reliability. *Educational and Psychological Measurement*, *59*(2), 234–247.

Rudner, L. M., & Liang, T. (2002). Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning and Assessment*, *1*(2), 1–22.

Slaughter, J. E., Christian, M. S., Podsakoff, N. P., Sinar, E. F., & Lievens, F. (2014). On the limitations of using situational judgment tests to measure interpersonal skills: The moderating influence of employee anger. *Personnel Psychology*, *67*(4), 847–885.

Süzen, N., Gorban, A. N., Levesley, J., & Mirkes, E. M. (2020). Automatic short answer grading and feed-back using text mining methods. *Procedia Computer Science*, *169*, 726–743.

Tavoosi, S. (2022). *Development and Validation of a Counterproductive Work Behavior Situational Judgment Test With an Open-ended Response Format: A Computerized Scoring Approach.* (Unpublished master's thesis). University of Central Florida.

Wang, Y., & Peng H.L. (2019). Validation on Automatic Scoring for Open-ended Questions in Chinese Oral Tests. *China Examinations*, *9*, 63–71.

[王妍, 彭恒利. (2019). 汉语口语开放性试题计算机自动评分的效度验证. *中国考试*, *9*, 63–71.]

Weekley, J. A., & Ployhart, R. E. (2005). Situational judgment: Antecedents and relationships with performance. *Human Performance*, *18*(1), 81–104.

Whetzel, D. L., & McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review, 19*(3)*,* 188–202.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2–13.

Xie, X. Q. (2013). Validation: From Reasonable to Plausible Interpretation of Test Score. *China Examinations*, *7*, 3–8.

[谢小庆. (2013). 效度: 从分数的合理解释到可接受解释. *中国考试*, *7*, 3–8.]

Xu, J. P. (2004). *Research on Teacher Competency Model and evaluation* (Unpublished doctorial dissertation). Beijing Normal University.

[徐建平. (2004). *教师胜任力模型与测评研究* (博士学位论文). 北京师范大学.]

Yang, L., Xin, T., Luo, F., Zhang, S., & Tian, X. (2022). Automated evaluation of the quality of ideas in compositions based on concept maps. *Natural Language Engineering*, *28*(4), 449–486.

Zhang, Y., Lin, C., & Chi, M. (2020). Going deeper: Automatic short-answer grading by combining student and question models. *User Modeling and User-Adapted Interaction*, *30*(1), 51–80.

Zhao, Y., Shen, Y., & Yao, J. (2019, August). Recurrent neural network for text classification with hierarchical multiscale dense connections. *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 5450–5456, Macao, PEOPLES R CHINA.

作者贡献声明：

徐　静：提出研究思路，设计研究方案；题目开发；数据采集、清洗和分析；论文起草；

骆　方：完善研究思路和方案；

马彦珍：模型实验对比；

胡路明：论文最终版本修订；

田雪涛：模型实验对比、论文最终版本修订。